Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 130 (2018) 464–471

Procedia Computer Science

www.elsevier.com/locate/procedia

9th International Conference on Ambient Systems, Networks and Technologies, ANT-2018 and the 8th International Conference on Sustainable Energy Information Technology, SEIT 2018, 8-11 May, 2018, Porto, Portugal

Filter hashtag context through an original data cleaning method

Didier Henry[a], Erick Stattner[a], Martine Collard[a]

[a]*LAMIA, University of French West Indies, 97110 Pointe A Pitre, Guadeloupe, France*

Abstract

Nowadays, social networks are one of the most used means of communication. For example, the social network Twitter has nearly 100 million active users who post about 500 million messages per day. Sharing information on this platform is unique because messages are limited in characters number. Faced with this limitation, users express themselves briefly and use sometimes a hashtag that summarizes the general idea of the message. Nevertheless, hashtags are noisy data because they do not respect any linguistic rule, may have several meanings, and their use is not under control. In this work, we tackle the problem of hashtag context which may have useful applications in several fields like information recommendation or information classification. In this paper, we propose an original data cleaning method to extract the most relevant neighbor hashtags of a hashtag. We test our method with a dataset containing hashtags related to several topics (such as sport, music, technology, etc.) in order to show the efficacy and the robustness of our approach.

© 2018 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the Conference Program Chairs.

Keywords: social media ; hashtag ; context ; data cleaning

1. Introduction

In just a few years, social media has become a huge platform for exchange and sharing that collects writing, opinions, thoughts and events of humanity. However, while these spaces of public exchange are today firmly anchored in our modern societies, we observe that ways of communicating have also evolved considerably through these media.

In particular, the limitation in terms of size imposed by certain platforms and the encouragement of rapid and impulsive sharing often incite users to share short messages, consisting of short words and accompanied by emoticons. Moreover, these messages are often associated with *hashtags*, that is to say one or several words preceded by the symbol '*#*', for instance: *#xboxone, #iphone, #tesla*.

Hashtags aim to join discussions on emergent topics [1], to identify or track conversations on the same event [2] or even to serve as the symbol of a community [3,4]. Thus, hashtags may be useful to find a related content on social media but do not respect any linguistic rule: *#musicislife, #yummm, #tlot*. In addition, a single hashtag may have several

* Corresponding author. Tel.: +590-059-048-3074 ; fax: +590-059-048-3086.
E-mail address: didier.henry@univ-antilles.fr

1877-0509 © 2018 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the Conference Program Chairs.
10.1016/j.procs.2018.04.050

meanings, for example, *#controller* can refer to a job or to an input device used for playing video games. Moreover, a hashtag present in a message may be inappropriate simply because their creation and sharing are not controlled. Therefore, the understanding of hashtags may be difficult for machines and humans because they are noisy data in terms of the lexicon, syntax and semantic.

In recent years, numerous researchers[5,6,7,8] have expressed interest in classification or category recognition of messages posted on social media. Michelson et al.[9] note that hashtags are ungrammatical and noisy. They have proposed a method linking a tweet to a tree of Wikipedia categories. Next this insight, Genc et al.[10] have shown that a Wikipedia-based technique produces better classification messages accuracy than Latent Semantic Analysis[11] (LSA) and String Edit Distance[12] (SED). In their work, Ferragina et al.[13] have focused on semantic hashtags classification. They have used a Hashtag-Entity Graph, where entities are linked to Wikipedia categories and a support vector machine classifier. Thus, we observe that Wikipedia has been used in several approaches and seems to be useful to both messages and hashtags classification task.

Messages and hashtags classification problem is closely linked to messages and hashtags recommendations. Li et al.[14] have proposed a method to suggest hashtags by using WordNet and a Euclidean similarity distance. In their approach, Godin et al.[15] have used a Latent Dirichlet Allocation (LDA) model to recommend hashtags from the message. Lu et al.[16] have introduced a recommendation tweets system based on user previous tweets. In a recent work, Gong et al.[17] have adopted a convolutional neural networks to perform the hashtag recommendation problem.

Messages classification may have useful applications for studies in the field of information diffusion. For instance, Romero et al.[18] have remarked that information propagation is different according to the topic. Myers et al.[19] have noticed that messages about education, art or work have a shorter reach than others. Several works[20,21] have observed discounts and promotions dissemination in social networks. Other researchers[22,23] have proposed diffusion models taking into consideration the message topic.

In this work, we are interested in the contextualisation of hashtags by other hashtags. We argue that such a context may improve information understanding both machines and humans, and may be useful in information recommendation/classification field. Nevertheless, some hashtags neighbour may be incoherent with the hashtag context. Our aim is to extract the most relevant hashtags related to a hashtag, so indirectly find messages dealing with close topic. To the best of our knowledge, not any work has focused on the hashtag context filtering.

The social network Twitter is a good case study to address this kind of problem. Indeed, Twitter has been a pioneer in the appearance of hashtags, and also contains a wide variety of users: professionals, individuals, politicians, associations, unions, companies, etc. In addition, this platform has nearly 100 million active users and 500 million messages are posted every day.

The rest of the article is organised as follows: Section 2 describes the proposed methodology. Section 3 details the experiments and the obtained results. Section 4 is dedicated to the discussion. Finally, Section 5 concludes and presents our future directions.

2. Methodology

Our aim is to filter the hashtag context which may contain irrelevant hashtags. We propose a data cleaning method based on the generation of three hashtags context: the *time context*, the *artificial context* and the *recent context*. In order to obtain the cleaned context, we choose hashtags present in at least two contexts (see Figure 1). Algorithm 1 presents this approach. Our method turns out to be language-independent as long as there are Wikipedia pages related several topics for the target language. In addition, our method is suitable for parallel computing.

Function `getCleanedContext(`*hashtag*`):`
  $HRecentList \longleftarrow getRecentContext(hashtag, nbTweets, nbRelatedHashtags, dateSince, dateUntil)$;
  $HTimeList \longleftarrow getTimeContext(hashtag, nbTweets, nbRelatedHashtags, dateSince1, dateUntil1, dateSince2, dateUntil2, dateSince3, dateUntil3)$;
  $HArtificialList \longleftarrow getArtificialContext(hashtag, nbTweets1, nbRelatedHashtags, nbTopicsWords, dateSince, dateUntil)$;
  $relatedHashtagsList \longleftarrow keepCommonHashtags(HRecentList, HTimeList, HArtificialList)$;
  return *relatedHashtagsList*;

Algorithm 1: Algorithm for cleaning the hashtag context

Download English Version:

https://daneshyari.com/en/article/6900014

Download Persian Version:

https://daneshyari.com/article/6900014

Daneshyari.com